PLoS GENETICS

Provisional DOI: 10.1371/journal.pgen.0020173.eor

Daniel A. Pollard, Venky N. Iyer, Alan M. Moses, Michael B. Eisen

Citation: Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Whole genome phylogeny of the *Drosophila melanogaster* species subgroup: Widespread discordance of mutations and genealogies with species tree. DOI: 10.1371/journal.pgen.0020173.eor

Future Article URL: http://dx.doi.org/10.1371/journal.pgen.0020173

Whole Genome Phylogeny of the *Drosophila melanogaster* Species Subgroup: Widespread Discordance with Species Tree & Evidence for Incomplete Lineage Sorting

Daniel A. Pollard[1], Venky N. Iyer[2†], Alan M. Moses[1†], Michael B. Eisen[1234]

1. Graduate Group in Biophysics, University of California, Berkeley, CA 94720, USA
2. Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA
3. Department of Genome Sciences, Genomics Division, Ernest Orlando Lawrence Berkeley National Lab, Berkeley, CA 94720, USA
4. Center for Integrative Genomics, University of California, Berkeley, CA 94720, USA
† These authors contributed equally to this manuscript.

Corresponding author:
Michael B. Eisen
1 Cyclotron Road, Mailstop 84-171
Berkeley, CA 94720
mbeisen@lbl.gov

Abreviations:

ABSTRACT

The phylogenetic relationship of the now fully sequenced species *Drosophila erecta* and *D. yakuba* with respect to the *D. melanogaster* species complex has been a subject of controversy. All three possible groupings of the species have been reported in the past, though recent multi-gene studies suggest that *D. erecta* and *D. yakuba* are sister species. Using the whole genomes of each of these species as well as the four other fully sequenced species in the subgenus *Sophophora*, we set out to investigate the placement of *D. erecta* and *D. yakuba* in the *D. melanogaster* species group and to understand the cause of the past incongruence. Though we find that the phylogeny grouping *D. erecta* and *D. yakuba* together is the best supported, we also find widespread incongruence in nucleotide and amino acid substitutions, insertions and deletions, and gene trees. The time inferred to span the two key speciation events is short enough that under the coalescent model, the incongruence could be the result of incomplete lineage sorting. Consistent with the lineage sorting hypothesis, substitutions supporting the same tree were spatially clustered. Support for the different trees was found to be linked to recombination such that adjacent genes support the same tree most often in regions of low recombination and substitutions supporting the same tree are most enriched roughly on the same scale as linkage disequilibrium, also consistent with lineage sorting. The incongruence was found to be statistically significant and robust to model and species choice. No systematic biases were found. We conclude that phylogenetic incongruence in the *D. melanogaster* species complex is the result, at least in part, of incomplete lineage sorting. Incomplete lineage sorting will likely cause phylogenetic incongruence in many comparative genomics data sets. Methods to infer the correct species tree, the history of every base in the genome and comparative methods that control for and/or utilize this information will be valuable advancements for the field of comparative genomics.

SYNOPSIS

To take full advantage of the growing number of genome sequences from different organisms, it is necessary to understand the evolutionary relationships (phylogeny) between organisms. Unfortunately, phylogenies inferred from individual genes often conflict, reflecting either poor inferences or real variation in the history of genes. In this study Pollard and colleagues examine relationships within the *Drosophila melanogaster* species subgroup, a group of flies with three fully-sequenced species whose phylogeny has been a source of controversy. Although the bulk of the data support a phylogeny with *Drosophila melanogaster* as an outgroup to sister species *Drosophila erecta* and *Drosophila yakuba*, large portions of their genes support alternative phylogenies. According to the authors, the most plausible explanation for these observations is that polymorphisms in the ancestral population were maintained during the two rapid speciation events that led to these species. Subsequent to speciation, polymorphisms were randomly fixed in each species and in some cases non-sister species fixed the same ancestral polymorphisms while sister species did not. In these cases the genes are correctly inferred to have conflicting phylogenies. The authors note that rapid speciation events will often lead to such conflict, which needs to be accounted for in evolutionary analyses.

INTRODUCTION

With the sequencing of twelve species from the genus *Drosophila*, the field of comparative genomics is now presented with the opportunity and challenge of understanding the function and history of every base in the model organism *Drosophila melanogaster* (*Dmel*). This process will hopefully result in the discovery of new biological phenomena and the development of new methodologies that will eventually help with the task of annotating other clades in the tree of life, particularly the human genome. Because most analyses of multiple genome sequences involve inferences about evolutionary history, they require an accurate description of the relationship of the species being analyzed.

The species history of the genus *Drosophila* has been the subject of numerous studies and the consensus from the literature suggests that the relationship of the twelve sequenced species is well resolved, with the exception of the species within the *Dmel* species subgroup and perhaps the placement of the Hawaiian, *D. grimshawi*, and the *virilis-repleta* species, *D. virilis* and *D. mojavenis* [1-5]. Within the *Dmel* species group, the placement of *D. erecta* (*Dere*) and *D. yakuba* (*Dyak*) relative to the *Dmel* lineage has been the subject of numerous conflicting studies [1-3,6-15]. Considering the placement of *Dmel*, *Dere* and *Dyak*, all three of the possible phylogenies (Figure 1) have received support. The topology (*Dmel*,(*Dere*,*Dyak*)), which we shall refer to as tree 1, was supported by studies of polytene chromosome banding sequences [6], satellite DNA [7], the *COI* and *COII* mitochondrial genes [3], mitochondrial DNA [16], the *fru* gene [17], the *Cu/Zn SOD* gene [18], the *H3* gene family [19], a concatenation of mitochondrial and nuclear genes [20], a concatenation of the genes *Adh*, *Adhr*, *Gld* and *ry* [8] and a concatenation of the genes *Adh*, *Amyrel*, *janA*, *janB* and *Sod* [9]. The topology ((*Dmel*,*Dere*),*Dyak*), which we shall refer to as tree 2, was supported by studies of an internal transcribed spacer region of ribosomal RNA genes [10], nucleotide sequences 5' of the *Amy* gene [15] and the *Adh* gene [8,21]. The topology ((*Dmel*,*Dyak*),*Dere*), which we shall refer to as tree 3, was supported by studies of protein electrophoresis [11], mitochondrial DNA [12], single copy nuclear and mitochondrial DNA hybridization [13], the *Adh* gene [1,14] and the *Amy* gene [15]. The support that each of these studies provides for the three phylogenies, however, is not uniformly strong. The most recent study by Ko et al using the concatenation of multiple nuclear genes provides the most compelling evidence, with 100% bootstrap support, for the placement of *Dere* and *Dyak* as sister taxa relative to the *Dmel* lineage. That Ko et al found such strong support for tree 1, despite using the *Adh* gene, which on its own has been found to support the other two trees, suggests that the past incongruence was likely the result of sampling variance [22,23]. Incongruence, however, can also be the result of numerous systematic biases [24-28] that are not overcome by increased sampling [29-31], as well as phylogenetically meaningful phenomena, such as lateral transfer [32] and incomplete linage sorting [25,33-48].

In this study, we set out to examine the possible causes of incongruence in this phylogeny and to investigate the placement of *Dere* and *Dyak* in the *Dmel* species subgroup, using the newly sequenced genomes in the genus *Drosophila*. Although we found that tree 1,

placing *Dere* and *Dyak* as sister species, is the best-supported tree, we found genome-wide incongruence in substitutions, indels and gene trees. We show that the branch separating the split of *Dmel* from the split of *Dere* and *Dyak* is sufficiently short that incomplete lineage sorting is a plausible explanation for the incongruence. We further show that the support for the three possible trees is non-randomly distributed across the genome such that adjacent genes supporting the same tree are more likely in regions of low recombination and substitutions supporting the same tree are most enriched roughly on the same scale as estimates of linkage disequilibrium, consistent with theoretical predictions under the coalescent [49]. We tested for obvious systematic biases and found that no factor we examined could account for the incongruence. We conclude by suggesting that incongruence due to incomplete lineage sorting has important implications for comparative genomics research.

RESULTS

Comparative Annotation of *Drosophila* Species

To analyze the phylogenetic history of the gene compliment of each of the seven fully sequenced species in the subgenus *Sophophora*, we mapped *Dmel* gene annotations onto each unannotated genome. 19,186 *Dmel* coding sequences were mapped to potential orthologous regions in each species using TBLASTN, and GeneWise was used to build gene models based on the *Dmel* gene in each region. These GeneWise models were matched back to *Dmel* translations using BLASTP and genes for which clear orthologs could be found were used in downstream analysis (see Methods). Peptide sequences from orthologs were aligned using TCoffee [50] and cDNA alignments were mapped onto the peptide alignments.

Species and Trees

Of these seven subgenus *Sophophora* species, we chose to use *Dmel*, *Dere*, *Dyak* and *D. ananassae* (*Dana*) for our initial analysis of the placement of *Dere* and *Dyak* within the *Dmel* species subgroup (we examine the effects of species choice on our results below). *Dmel* was chosen because the annotations were mapped from *Dmel* and it is the primary model organism of the subgenus. *Dsim* and *Dsec* were excluded from initial analysis because they were assumed to provide mostly redundant information to *Dmel* and they reduced the number of clear orthologs spanning the species by 2544 genes, presumably because of lower sequence coverage and issues regarding the assembly of polymorphic reads in *Dsim*. *Dana* was chosen over *Dpse*, because it is the closest fully sequenced outgroup to the *Dmel* species subgroup. 9405 genes were found to have clear orthologs in all four of the chosen species. Figure 1 shows the three possible unrooted trees relating the species.

Genome Wide Incongruence

We began our analysis looking directly at the genome-wide counts of amino acid substitutions, nucleotide substitutions and insertion/deletion (indel) events that were

informative with respect to each of the three possible trees (see Methods). For all three characters, tree 1, which groups *Dere* and *Dyak* together, was found to have the most support (Figure 2ABC). By a majority-rule consensus, tree 1 would be inferred to be the species tree, consistent with the findings of Ko et al [8]. The high proportion of substitutions and indels supporting the alternate trees, however, suggests a poorly resolved tree and pervasive incongruence.

What is the cause of this incongruence? The incongruent substitutions could be the product of any of a number of systematic biases but the incongruent indels are unambiguous characters that are more difficult to explain as methodological artifacts [51,52]. The population genetic theory of the coalescent states that sufficiently close speciation events will lead to incongruence due to incomplete lineage sorting (Figure 3) [38]. Below we explore the compatibility of our data with the coalescent as well as test for possible systematic biases.

Maximum Likelihood Gene Trees Show Incongruence

We first repeated our analysis using maximum likelihood (ML) methods [53,54] to measure the informative divergence spanning the inferred speciation events and test the robustness of the incongruent substitutions using more complex models of sequence evolution. ML analysis is not currently scalable to entire genomes in a single calculation, so we partitioned the genome into individual genes. If incomplete lineage sorting is the underlying cause of the incongruence, such a partition might also reveal variation in allelic histories that multi-gene concatenations could obscure [27,45,55]. Wanting to capture both the observed nucleotide and amino acid differences across the species [56], we used the F3x4 codon-based model from the PAML package [57] to compare the likelihood of each tree given each cDNA alignment (we test other models below). Consistent with the parsimony-based analysis, the majority of genes (57.8%) support tree 1, while a high proportion (42.2%) support the other two trees (Figure 2D).

The median synonymous divergence trees for the sets of genes supporting each tree are: (dmel:0.1301,(dere:0.1095,dyak:0.1201):0.0664,dana:1.3246) for tree 1, ((dmel:0.1744,dere:0.1076):0.0498,dyak:0.0757,dana:1.2871) for tree 2 and ((dmel:0.1801,dyak:0.1163):0.0454,dere:0.0719,dana:1.3147) for tree 3 (Figure 4). The branches between the speciation events are quite short, with the tree 1 branch being the longest at only 0.066, suggesting that these species split in rapid succession.

Incongruence Is Expected For These Species Under the Coalescent.

Is the time spanning these speciation events short enough to expect the observed levels of incongruence? Using the coalescent, the probability of congruence, or monophyly, can be directly calculated for the three taxa case using the equation p(congruence)=1-2/3exp(-*t*), where *t* is the time between speciation events in units of generations/2Ne and Ne is the effective population size [58-60]. Figure 5 shows this probability graphically as a function of *t*. In order to go from an estimate of the informative divergence to this probability, the substitutions per site per year, the ancestral generation time and the

ancestral population size must be known. Synonymous-substitutions per site per year has been estimated to be in the range of $1\text{-}2\text{x}10^{-8}$ in *Drosophila* [1,13,61,62]. Generations per year for the extant taxa in the *Dmel* species subgroup is about 10 and can be used as an estimate for the ancestral generation time [63]. The ancestral population size has been estimated in the range of 10^6 to 10^7 but this should be considered a poorly resolved parameter [64]. Theoretically, the median informative branch length measured above includes both divergence prior to the first speciation event and divergence between the two speciation events. If we take the informative divergence estimated from genes supporting the alternative trees to represent the expected amount of divergence prior to the first speciation event (0.05 and 0.045 for trees 2 and 3 respectively) and subtract their average (0.0475) from the tree 1 total informative divergence (0.066), we can get an estimate of the informative divergence spanning the two speciation events (0.019). This leads to an estimate of $9.5\text{x}10^5$ to $1.9\text{x}10^6$ years or $9.5\text{x}10^6$ to $1.9\text{x}10^7$ generations. The range of values for t becomes 0.48 to 9.5, which produces probabilities for congruence in the range of 0.59 to 0.99995 (Figure 5). Although the uncertainty in these parameter estimates does not permit us to say that incongruence would be guaranteed, they do allow us to say that incongruence due to incomplete lineage sorting is expected under plausible assumptions about these species' ancestral population and speciation events.

Spatial Structure of Tree Support

Given that we observed incongruence in individual sites as well as for whole genes, we wanted to better understand the extent to which sites supporting the same tree are spatially correlated, with a particular interest in the compatibility of this structure with the incomplete lineage sorting hypothesis. The above analysis of gene trees suggests that sites can be correlated out to the length of genes. To see if this correlation extends beyond individual genes we looked for blocks of adjacent genes supporting the same gene and tested for unusual block lengths. Using permutations of ML gene tree state to obtain significance, we found gene tree block lengths at expected frequencies with the exception of an excess of long blocks supporting tree 3, in the range of 250kb to 700kb, three of which were highly significant ($p<0.05$).

If the blocks of genes supporting the same tree were the product of incomplete lineage sorting, then regions of low recombination ought to have larger blocks [65]. Although the ancestral recombination rates are not known, we looked to see if block lengths are correlated with *Dmel* recombination rates [66]. We found a weak negative correlation for all blocks (Pearson's $R=-0.13, p<0.1$) as well for blocks for each specific tree, with tree 2 blocks showing the strongest correlation (Pearson's $R=-0.30, p<0.05$). These weak correlations suggest a minor role for recombination rates in determining the spatial structure of support for different trees across the genome, however, there are many reasons for why strong correlations would not be expected, including poorly conserved recombination rates across these species [67-69] and gene conversion in regions of low recombination [70-72]. Nonetheless, these weak correlations establish a connection between recombination and the spatial structure of support that is at least consistent with lineage sorting. We next looked at the spatial correlation of individual sites to understand the spatial correlation at a finer scale.

Using the whole genome frequencies of informative amino acid and nucleotide substitutions supporting each tree, we looked to see if sites supporting the same tree are locally enriched across chromosomes (see Methods for more details). Figure 6 shows that informative amino acid and nucleotide substitutions supporting the same tree cluster together on the scale of less than 8kb for trees 1 and 2 and less than 2kb for tree 3. These local deviations in the frequencies of informative substitutions from the expected frequencies are quite highly significant (X^2 test, $p<10^{-10}$).

What forces might have shaped these clusters of informative sites supporting the same tree? Under the coalescent, linked neutrally evolving sites supporting the same tree have been proposed to be correlated at an expected distance equal to linkage disequilibrium [49]. Linkage disequilibrium in *Dmel* has been estimated to extend to the length of a few kb [73], suggesting that our results are consistent with theoretical expectations [49]. Theoretical considerations together with recent empirical evidence from *Dmel*, however, imply that neutral sites would not be expected to be in disequilibrium at distances greater than a few hundred base pairs [74,75], suggesting that perhaps selection has acted to increase the scale of these correlations [65]. Regardless of the influence of selection, the structure of the support for different trees across the genome is consistent with recombination acting within the context of incomplete lineage sorting.

Additional support for this conclusion comes from the observation that mitochondrial genes exhibit no incongruence (Montooth K. & Rand D., Personal Communication). This is expected, as recombination is not thought to occur in the mitochondrial genome. While mitochondrial evolution differs from nuclear evolution in more ways than just recombination [76], the complete lack of incongruence is nevertheless striking.

Thus far we have presented results suggesting that incomplete lineage sorting is a plausible explanation for the observed incongruence. We next sought to rule out alternate explanations.

Statistical Support For Incongruence

Is the incongruence in gene trees unexpected given the strength of support for each inference? To address this question, we used the bootstrap [77] value, RELL [78], from 10,000 replicates as an estimate of the expected incongruence due to chance alone. Taylor and Piel have shown for a large set of yeast genes, originally reported by Rokas et al [79], that there is no significant difference between nonparametric bootstrap values and accuracy, as measured by congruence [80]. Earlier work suggests that bootstrap values are conservative and likely to underestimate accuracy [81,82]. Figure 7A shows the proportion of genes supporting each tree in bins of bootstrap value. Unlike the yeast phylogeny, our observed incongruence consistently exceeds that expected by bootstrap values. Thus the incongruence for these four species using the F3x4 codon model appears to be statistically significant.

Incongruence Is Robust To Model Choice

We next tested if the incongruence is robust to model choice. An empirical study of model choice and accuracy by Ren et al found that codon based models are able to recover both recent and deep divergences well, while nucleotide based models are less efficient at deep divergences and amino acid based models are less efficient at recent divergences [56]. They also found that while more complex models fit the data better, they are not necessarily more accurate, a conclusion that has been made by other studies [83,84]. We looked at six models: nucleotide (HKY,HKY+G), codon (F3x4,F3x4+G) and amino acid (WAG+F,WAG+F+G) based models both with and without a discrete gamma model of variable rates amongst sites (see Methods). Incongruence was found to exceed expected levels from bootstrap values across all models, suggesting that the incongruence is indeed robust to model choice (Figure S1).

Comparing congruence across models, simpler models seem to produce more congruence than more complex models (Table 1). For each of the three types of models, addition of a discrete gamma resulted in lower congruence. For the models without discrete gamma, HKY was more congruent than F3x4, which was more congruent than WAG+F, perhaps due to the relatively recent divergences in this phylogeny. Interestingly, the more complex models, F3x4+G for nucleotides and WAG+F+G for amino acids, fit the alignments better for most genes, according to Akaike's information criterion (Table 1) [85]. Thus, consistent with the finding of Ren et al with the yeast dataset [56], more complex models fit the data better but produce less congruence.

Species Choice Does Not Explain The Observed Incongruence

To evaluate the robustness of the incongruence to species choice we examined the set of 5778 genes for which a clear ortholog could be found in all seven fully sequenced species in the subgenous *Sophophora*: *Dmel*, *Dsim*, *Dsec*, *Dere*, *Dyak*, *Dana* and *Dpse*. All 21 possible species combinations that include *Dere* and *Dyak* and at least one of *Dmel*, *Dsim* and *Dsec* as well as at least one of *Dana* and *Dpse* were considered. The HKY model was used both because it was found to produce the most congruence in the original four species as well as because it is considerably more computationally efficient than the codon models. Across all species combinations, incongruence is consistently greater than expected from bootstrap values, suggesting that incongruence is not species choice dependent (Figures S1A & S2).

Ranking species combinations by levels of congruence reveals that our original species choice produces the most congruence (Table 2), suggesting that our estimates are conservative. The relative congruence of the species combinations appears non-random, with respect to presence or absence of individual species, so we calculated the average congruence for each species across the combinations containing that species. Although the average congruence is very similar for each species, we found that *Dana* (82.4%) contributes most to congruence, while *Dsim* (80.8%), *Dsec* (80.4%) and *Dpse* (79.7%) contribute roughly equally and *Dmel* (78.9%) actually contributes least to congruence. We note that the presence of *Dmel* in the most congruent species combination goes

against this general trend, perhaps reflecting further complexities in the impact of species choice on congruence.

Consistency

Although the incongruence appears to be robust to model and species choice, a much more stringent test is to look at incongruence in the partition of genes that consistently support the same tree across all models and across all species combinations [86]. Of the 5778 genes analyzed, 2347 are consistent across all models and of those, 1600 (68.2%) are congruent while 443 (18.9%) support tree 2 and 304 (12.9%) support tree 3. Similarly, 1918 genes are consistent across species combinations and of those, 1474 (76.8%) are congruent while 291 (15.2%) support tree 2 and 153 (8%) support tree 3. Finally, 970 genes are consistent across all models and all species combinations and of those, 804 (82.9%) are congruent while 101 (10.4%) support tree 2 and 61 (6.3%) support tree 3. This conservative partitioning reduces the amount of incongruence but does not eliminate it. We note that under the incomplete lineage sorting hypothesis, incongruent genes are expected to have accumulated fewer informative substitutions (Figure 4) and therefore might be expected to be less robust to such a consistency test.

To assess the statistical significance of the incongruence in the partition of genes consistent across all models and species combinations [31], we used the HKY model bootstrap values from the *Dmel*, *Dere*, *Dyak* and *Dana* species combination to look at congruence as a function of bootstrap value. As shown in figure 7B, the congruence is less than expected for the highest bootstrap values. For the 521 genes with bootstrap values between 0.9 and 1.0, which is more than half of consistent genes, the incongruence was highly significant (X^2 test, $p<10^{-3}$).

To further test whether the statistical support from the incongruent genes is the result of consistent signal, as opposed to having hidden support [87] for tree 1, we concatenated the 804 consistent tree 1 genes, 101 consistent tree 2 genes and 61 tree 3 genes into three large alignments and repeated the ML analysis for the *Dmel*, *Dere*, *Dyak* and *Dana* species combination and the HKY model. Interestingly, each tree-specific concatenation supported its tree with 100% bootstrap support [88]. Thus the signal for incongruence appears to be consistent, highly significant and robust to model and species choice consistency partitioning.

Sequence & Evolutionary Properties

We next looked at sequence and evolutionary properties of the genes supporting each tree to see if any clear biases could explain the incongruence. The properties we examined are sequence quality, gene length (measured in ungapped codons in the alignment), base composition (GC content) across the species at each position in the codon, transition:transversion ratio (kappa), dN/dS, informative synonymous divergence (ISD), ratio of informative synonymous divergence to non-informative synonymous divergence (RINSD), and total synonymous divergence (TSD). Table S1 shows the correlation of bootstrap values to each of these properties for the whole set of genes, genes supporting

each tree, the set of genes found to be consistent across models and species combinations, the genes that consistently supported each tree as well as the set of inconsistent genes. Distributions for each property are shown in Figures 8 and S3-8.

The strongest and most consistent correlations with bootstrap value are for ISD and RINSD (Table S1), which are in essence the signal and signal to noise. We've already shown that the median informative divergence in the genes supporting tree 1 is greater than that for the genes supporting trees 2 and 3 (Figure 4). Reflecting this, the distributions of ISD and RINSD for genes supporting trees 2 and 3 are shifted toward lower values compared to genes supporting tree 1 (figures 8A & S3A). Comparing consistent genes and inconsistent genes reveals that nearly all genes with ISD values close to zero are classified as inconsistent (figure 8B). Amongst consistent genes, those supporting trees 2 and 3 still have distributions of ISD and RINSD shifted slightly toward lower values compared to those supporting tree 1 (figures 8C & S3B). The fact that incongruent genes are expected to have lower ISD than congruent genes under the incomplete lineage sorting model (see above) and the fact the ISD and RINSD distributions are highly overlapping for each of the three trees suggests that a simple lack signal or low signal to noise cannot explain the observed incongruence.

The long branch out to *Dana* (figure 4) presents the concern that the incongruence may be due to homoplasy and perhaps long branch attraction. TSD is distributed nearly identically across all sets of genes, including consistent and inconsistent genes, with a very slight bias toward tree 2 and 3 genes and inconsistent genes having lower TSD (figures 8D & S4). Although this does not rule out homoplasy as a source for noise in the inference of gene trees, it appears that regions with high mutational rates are not biased toward supporting incongruent or inconsistent genes [89], making it a less likely explanatory factor. Additionally, although the trees in figure 4 are not ultrametric (leaves equidistant from internal nodes), they are biased in the opposite direction as would be expected under long branch attraction, with the shortest branch in the *Dmel* species subgroup pairing with the longest branch out to *Dana*. Thus, homoplasy and long branch attraction do not appear to be responsible for the incongruence.

Another possibility is that sampling variance in short genes is leading to the incongruence [90]. We've already shown that a concatenation of the consistent genes supporting each tree gives 100% bootstrap support, making sampling variance an unlikely explanation. Gene length is very similar across the sets of genes supporting each tree but tree 1 genes tend to be slightly longer than genes supporting trees 2 and 3 (figure 8E). Gene length is also weakly correlated with bootstrap value for the whole set, consistent genes and tree 1 genes (both inconsistent and consistent) (Table S1). Our above results on the spatial correlation of sites, however, suggests that genes that extend more than a few kb would not be expected to be enriched for sites supporting the same tree above their background frequencies. We also found that enrichment is most pronounced for tree 1 sites and less so for incongruent sites. This increased mosaic structure [91] in incongruent genes is likely to be responsible for most of the shift to slightly larger genes in the tree 1 genes. The influence of sampling variance, however, is reflected in the shift of inconsistent genes compared to consistent genes toward shorter lengths. Thus the small decrease in

long genes in the incongruent set is probably a result of the spatial clustering of sites while the small increase in short genes may be a combination of that effect and noise from sampling variance. Regardless, gene length is so similar across trees that it is unlikely to explain the incongruence.

GC content has been estimated to vary considerably across the species in the *Dmel* species subgroup [92] and is therefore a major concern for systematic bias. We found that GC content is highly similar across species at 1st and 2nd codon positions but varied systematically at the 3rd codon position (figures 8F & S5AB). *Dmel* and *Dana* have nearly identical distributions of 3rd codon position GC content, which is shifted toward lower values compared to *Dere* and *Dyak*, which also have nearly identical distributions. This bias in GC content across species is very conservative with respect to the inference of incongruent genes because the incongruence would need to overcome the signal from base composition alone [29]. To further verify that this bias only works to decrease the incongruence, we converted the cDNA alignments into R's and Y's, for purines and pyrimadines respectively, and repeated the ML analysis using the F81 model of evolution, effectively averaging the contribution of GC and AT content and only measuring transversions [29,93,94]. As expected, incongruence actually increases (45.2%) under the RY-coding and is still statistically significant (figure S9). Other methods, for example those of Galtier and Gouy [95,96] and Gu and Li [97], attempt to explicitly model non-stationary evolution, rather than control for it. These methods might reveal more precisely the underestimation of incongruence due to the base composition bias in these species but are not expected to provide an explanation for the observed incongruence.

Sequence qualities, transition:transversion ratios and dN/dS values were found be distributed similarly across trees, suggesting they are unlikely factors for systematic bias (figures S6-8).

Sequence Properties Associated With Spatial Clustering

We last look to see if the spatial clustering of sites supporting the same tree could be explained by evolutionary rate or base composition variation. To examine the relationship of evolutionary rate and the clustering of sites supporting each tree, we measured total divergence and the fraction of sites supporting each tree in overlapping windows across the chromosomes. For windows of size 5kb or 1kb no correlation can be found between divergence and the fraction of sites supporting each tree, suggesting that evolutionary rate is unlikely to explain the spatial clustering. To test if changes in GC content could explain the clustering of sites we used the RY-coded alignments (described above) [29,93,94] and repeated the spatial clustering analysis. Figure S10 shows that sites are still correlated in a similar range of a few kb, suggesting that variance in GC content is unlikely to be causing the spatial clustering of sites. Thus both the incongruence as well as the spatial clustering of sites appears to be robust to the sequence and evolutionary properties examined.

DISCUSSION

We initially set out to confirm the placement of *Dere* and *Dyak* as sister species, relative to the *Dmel* lineage, in the *Dmel* species subgroup, using the fully sequenced genomes of seven species in the subgenus *Sophophora*. Although we did find that the best-supported phylogeny is that which places *Dere* and *Dyak* as sister species, we also found pervasive incongruence of substitutions, indels and gene trees (figure 2). While incongruence in substitutions and gene trees could be the result of systematic biases, the incongruent indels, particularly unique insertions, presented strong enough evidence for unbiased incongruence that we also considered incomplete lineage sorting as a possible explanation. Assuming plausible values of substitution rate, generation time and ancestral population size, we found that the time between the split of *Dmel* and the split of *Dere* and *Dyak* is sufficiently short that incomplete lineage sorting would be expected (figures 3-5). Interestingly, we observed that the support for each of the three trees has a spatial structure across the genome, which is related to low recombination, both locally and globally (figure 6). This further supports the hypothesis that the observed incongruence is due, at least in part, to incomplete lineage sorting.

To test for other plausible explanations we examined model choice, species choice and variation in sequence and evolutionary properties and found no obvious candidate factors to explain the incongruence or the spatial structure of support for trees (tables 1 & 2; figures 7, 8 & S1-10). We therefore conclude that incomplete lineage sorting is the best going explanation for the lack of resolution in this phylogeny.

Nevertheless, we likely did not exhaust the possible tests for alternate hypotheses for incongruence and suspect that this dataset will prove an interesting area for systematic research, much as the Rokas et al yeast dataset has [69]. Comparing our results to the yeast dataset reveals important differences: there is significant incongruence beyond what would be expected by chance (figure 7A), the level of incongruence is relatively robust to model choice (tables 1 & 2; figures 7B & S1), and basic sequence properties, like GC content, vary in ways that are conservative with respect to the incongruence (figures 8 & S3-10) [29]. Similar to the yeast dataset, however, we find that the evolutionary model that maximizes the congruence (or accuracy as Ren et al refer to it) is typically the simplest (HKY) while the model that fits the data best is the most complex (F3x4+G) (table 1) [56].

To further understand the extent and nature of incomplete lineage sorting in the *Dmel* species subgroup, we suggest several types of future studies. First, to further test the agreement of the observed incongruence with theoretical predictions, better estimates of the ancestral effective population size, mutation rates, time between speciation events, ancestral recombination events [98] and examining the effects of selection (both directional and balancing [99]) would be of clear benefit. In addition, of great interest will be studies of lineage sorting across all taxa in the species group (especially the *Dsim* species complex [39]) and the influence of migration and gene flow on the symmetry of lineage sorting (because tree 2 is asymmetrically favored). Genome-wide population data already exists for *Dsim* and is expected for *Dmel*, which have the potential to help in the

effort to understand these processes. Finally, methodological improvements might include increased large-scale taxon sampling, particularly from closely related taxa outside the species subgroup, such as the *D. suzukii* and *D. takahashii* subgroups [3], would alleviate potential biases introduced by the long branches out to *Dana* and *Dpse*.

Although this study should prove quite valuable to the increasing numbers of comparative genomics researchers studying the genus *Drosophila*, we believe our findings have important implications for comparative genomics as a whole. The idea that speciation events have occurred in rapid bursts throughout the tree of life [100-102] is likely broadly understood (for example the short branch connecting the human, mouse and dog lineages [103]), but the idea that genomes may be mosaics of conflicting genealogies as a result of rapid speciation is perhaps less well appreciated. As more species are sequenced, particularly the dense taxon sampling that is currently beginning in model organism clades, increasing numbers of close speciation events will likely result in many cases of incomplete lineage sorting in genome-scale data. As many methods used in comparative genomics require an accurate phylogeny, the comparative genomics community must develop methods that are robust to or take into account variation in phylogeny.

We envision three types of methods that will need to be developed to appropriately account for this kind of variation. The first are methods that can infer the most likely species tree using an entire genome in a single calculation, considering lineage sorting explicitly. The second are methods that can infer the most likely history of every base in every species, given the species tree. Lastly, comparative genomics methods that use phylogenies would need to be altered to control for and utilize the output from the second kind of method. Progress is being made in the first two categories [27,38,47,48,98,104-113] although no currently available method can deal with a whole genome dataset such as this one. Though well appreciated in the systematics and population genetics communities, the issue of incomplete lineage sorting is rarely considered in the bioinformatics and comparative genomics communities, so the third category of method is virtually non-existent. Accounting for variation in evolutionary histories will have different effects on different classes of methods, but we suggest that parsimony-based methods would be most strongly affected. An important example of such a phylogeny-based method is genome-wide multiple alignment using a guide tree (i.e. [114] & [115]), which is the first step in nearly all comparative genomic analyses. The availability of genome-scale datasets such as the one analyzed here should allow rapid progress in all three of these types of methods; we suggest that their development will be of great benefit to the evolutionary and comparative genomics community in the near future.

METHODS

Assemblies

Dmel release 4.2 genome, cDNA and translation sequences were downloaded from Flybase (http://www.flybase.net). Pre-publication assemblies for *Dere* and *Dana* (dated August 01st, 2005), sequenced and assembled by Agencourt Bioscience, and for Dsec

(dated October 28th, 2005), assembled and sequenced by the Broad Institute were downloaded from the Berkeley AAA website (http://rana.lbl.gov/drosophila/). The pre-publication assemblies for *Dyak* (dated July 4th, 2004) and Dsim (dated June 2nd 2005) were downloaded from the Washington University School of Medicine Genome Sequencing Center's website (ftp://genome.wustl.edu/pub/seqmgr/yakuba/). The Dpse v1.04 assembly was downloaded from Flybase.*Dere, Dyak* and *Dana* assemblies can be found in *Dsim_Assembly.fasta, Dere_Assembly.fasta*, *Dsec_Assembly.fasta, Dyak_Assembly.fasta, Dpse_Assembly.fasta* and *Dana_Assembly.fasta*, which are published as supporting information. Sequencing traces corresponding to these genomes are in the NCBI trace archive (http://ncbi.nlm.nih.gov/Traces/trace.cgi, *species_code*, 'DROSOPHILA ERECTA', 'DROSOPHILA YAKUBA', 'DROSOPHILA ANANASSAE', 'DROSOPHILA SIMULANS', 'DROSOPHILA SECHELLIA', 'DROSOPHILA PSEUDOOBSCURA').

Comparative Annotation

Each of the sequence assemblies were annotated separately by mapping *Dmel* gene models onto the unannotated genome in a pairwise fashion using a modified reciprocal-BLAST approach [116] to assign orthology/paralogy relationships, and a comparative gene finder, GeneWise [117,118], to build gene models. The annotation pipeline consisted of three steps: (I) For each *Dmel* translation, we used the protein sequence as a NCBI TBLASTN [119] query (e-value threshold 1e-3) against the scaffolds of the target assembly. (II) The scaffolds were ordered by the hit e-value reported by TBLASTN and up to two regions were selected from the two best scaffolds and used as input to construct gene models using GeneWise. To improve the chance of constructing a complete gene model using GeneWise, the regions were selected by clustering HSPs on the scaffold such that every HSP within 100kb of another HSP was included in the same region, and a buffer of 10kb was included at the ends of the regions. (III) The predicted translations of the models reported by GeneWise were then used as BLASTP queries against a database of *Dmel* translations, with an e-value threshold of 1e-3.

We then assigned orthology/paralogy relationships using a heuristic algorithm that takes into account (a) the rank of the starting *Dmel* translation in the BLASTP results, (b) the rank of alternative translations from the gene corresponding to the starting *Dmel* translation, and (c) whether or not there were highly ranked hits to genes other than the gene corresponding to the starting *Dmel* translation. One-to-one orthology was assigned when the only top-ranked hits in the BLASTP results were translations from the gene corresponding to the starting *Dmel* translation. Hits having e-values within one order of magnitude were considered to be equivalently ranked. For genes with more than one translation with clear orthologs in each species, the first historically annotated (translation with the lowest letter ID) was used to represent the gene.

cDNA and translation sequences can be found in *Dsim_cDNAs.fasta*, *Dsim_translations.fasta*, *Dsec_cDNAs.fasta*, *Dsec_translations.fasta*, *Dere_cDNAs.fasta*, *Dere_translations.fasta*, *Dyak_cDNAs.fasta*, *Dyak_translations.fasta*,

Dana_cDNAs.fasta, *Dana_translations.fasta*, *Dpse_cDNAs.fasta* and *Dpse_translations.fasta*, published as supporting information.

Informative Substitutions & Indels

Informative substitutions supporting each tree were counted across all cDNA and peptide alignments. Only single substitutions that split the four species into two groups of two were considered. Informative substitutions for tree 1 grouped *Dmel* and *Dana* together and *Dere* and *Dyak* together. Likewise tree 2 grouped *Dmel* and *Dere* together and tree 3 grouped *Dmel* and *Dyak* together.

Informative indels supporting each tree were counted across all peptide alignments. Indels were classified as informative in the same way substitutions were. Indels were further filtered to avoid artifacts from alignment errors. Only indels with five amino acids of perfect identity in flanking sequences, with no mono-, di- or tri-amino acid repeats, were included. Insertions were inferred based on an absence in *Dana* and one of the ingroup species. Such insertions, where the inserted sequence is the same in the two species containing it, provided strong, unambiguous characters.

ML Gene Trees

The Codeml program of the PAML package (version 3.14) [57,120] was run on each gene using the following three unrooted trees: Tree1 - ((*Dmel*,(*Dere*,*Dyak*),*Dana*), Tree2 - ((*Dmel*,*Dere*),*Dyak*,*Dana*) & Tree3 - ((*Dmel*,*Dyak*),*Dere*,*Dana*) (see figure 1). Codeml was run using the F3x4 model, such that equilibrium codon frequencies were calculated from the average nucleotide frequencies at the three codon positions (CodonFreq = 2), amino amino acid distances were equal (aaDist = 0), one dN/dS value was estimated for all lineages using an initial value of 0.4 (model = 0, fix_omega = 0, omega = 0.4), the transition:transversion ratio was estimated with an initial value of 2 (fix_kappa = 0, kappa = 2), substitution rates across sites were set to be equal (fix_alpha = 1, alpha = 0), substitution rates were allowed to vary freely across lineages (clock = 0) and codons with ambiguous positions (gaps or Ns) were ignored (cleandata = 1).

Spatial Analysis

Based on the maximum likelihood tree for each gene, the genome was divided up into blocks supporting each tree. A ten-gene sliding-window was used to calculate a running average of the support for each tree along each chromosome. Each window was assigned a tree based on the most frequent genealogy in the window. Each gene was then reassigned a tree based on the most frequent tree of all the windows that contained it. This effectively allows the neighbors of a gene to influence its assignment, and near neighbors have more influence than far neighbors. Adjactent genes which support the same tree were combined together into blocks. To measure the significance of the size of the blocks, the labels for each gene in the genome were randomized 1000 times and the blocks were recalculated for each replicate, using the windowing method described above. Recombination rates for a subset of genes in *Dmel,* calculated by Hey and Kliman

[66] using the *R* statistic, were downloaded. The average *R* in each block was calculated where a gene could be found in their set. The Pearson correlation of the average *R* within blocks and the length of blocks was calculated using the R statistics package.

Informative substitutions in genes were used to look at the structure of support for the different trees across the genome independent of the likelihood inference. The counts of each type of informative substitution were calculated in 60 non-overlapping 1kb windows surrounding each informative substitution across all chromosomes. The frequency of each kind of informative substitution across the whole genome was used to calculate an expected count for each 1kb window. In each window, the enrichment of informative substitutions supporting the same tree was calculated. The X^2 significance of windows was calculated by comparing the observed frequencies of informative mutations supporting each tree with the genome averages of those frequencies.

Bootstrap Values

RELL bootstrap values [78] from 10,000 replicates were taken from the Codeml output.

PAML Models

All models were run using the same settings as described above for F3x4 except where HKY (model = 4) or WAG+F (model = 3) was specified and where the gamma function was used (fix_alpha = 0, alpha = 1.0, ncatg = 8).

AIC

AIC was calculated as AIC = -2 ln L + 2 N, where L is the likelihood of the model given the data and N is the degrees of freedom [85]. Only consistent genes were used in this analysis so the tree was the same across all models. The likelihood and degrees of freedom were taken directly from PAML output. HKY, HKY+G, F3x4 and F3x4+G were compared and WAG+F and WAG+F+G were compared.

Sequence & Evolutionary Properties Analysis

The sequence quality in each species was calculated as the mean sequence quality score of the coding bases. Bootstrap value, length, GC content, transition:transversion ratio, dN/dS, informative synonymous divergence, non-informative synonymous divergence and total synonymous divergence were taken directly from the PAML output for the ML tree from the original analysis using the F3x4 model and the *Dmel*, *Dere*, *Dyak* & *Dana* species combination. The Spearman rank correlations were calculated using the R statistics package [121].

Divergence Windows

To examine the correlation of divergence with the proportion of sites supporting each tree in local areas across the genome we used 5kb and 1kb windows, overlapping by 2.5kb

and 0.5 kb respectively. Using the synonymous site divergences reported by Codeml from the original analysis, we calculated the synonymous divergence per coding site in each window. We also calculated the proportion of sites supporting each tree in each window. Windows with no synonymous coding sites were excluded.

Acknowledgements

We thank Agencourt Bioscience, the Broad Institute and Washington University School of Medicine Genome Sequencing Center for pre-publication access to the genome sequence data. We thank Hiroshi Akashi and two anonymous reviewers for critical reading of the manuscript and helpful suggestions. We thank Peter Andolfatto, Doris Bachtrog, John Novembre, Joshua Pollack and Montgomery Slatkin for discussions regarding the coalescent and the spatial correlation of substitutions. We thank Matt Hahn and Alisha Holloway for advice on phylogenetics. We thank Angela Depace, Justin Fay, Hunter Fraser, Emily Hare, Suzanne Lee, Richard Lusk, Stewart McArthur, John Novembre, Joshua Pollack, Montgomery Slatkin, Erica Rosenblum and Jody Westbrook for comments on the manuscript.

Author Contributions

DAP designed the research, performed the research, analyzed the data and wrote the paper. VNI contributed to the research design, the comparative annotations and the gene blocks analysis. AMM contributed to the research design and the coalescent analysis. MBE contributed to the research design. All authors contributed to the writing of the paper.

Figure legends

Figure 1. Phylogenies
The three possible phylogenies for *D. melanogaster* (Dmel), *D. erecta* (Dere) and *D. yakuba* (Dyak) with *D. ananasae* (Dana) as an outgroup.

Figure 2. Widespread incongruence of substitutions, indels and genes trees
A. The proportion of informative nucleotide substitutions in 9405 genes supporting each of the three trees. 170,002 (44.7%) nucleotide changes support tree 1 (red), 112,278 (29.5%) support tree 2 (green) and 98,117 (25.8%) support tree 3 (purple). B. The proportion of informative amino acid substitutions in 9405 genes supporting each of the three trees. 28,628 (49.3%) amino acid changes support tree 1 (red), 15,182 (26.2%) support tree 2 (green) and 14,203 (24.5%) support tree 3 (purple). C. The proportion of informative insertions or deletions (indels) in 9405 genes supporting each of the three genes. Indels were filtered requiring five flanking amino acids of perfect identity and no repetitive sequence. 2 deletions and 6 insertions (66.7%) support tree 1 (red), 1 deletion and 1 insertion (16.7%) support tree 2 (green) and 2 insertions (16.7%) support tree 3 (purple). Similar proportions but much larger counts are found when the indels are not filtered. D. The proportion of 9315 genes with maximum likelihood support for each of

the three trees. 5381 (57.8%) support tree 1 (red), 2188 (23.5%) support tree 2 (green) and 1746 (18.7%) support tree 3 (purple).

Figure 3. Incomplete lineage sorting
The history of a gene (colored lines) is drawn in the context of a species tree (grey bars). New lineages arising from new polymorphisms in the gene are drawn in different colors. In this case, the two alleles in the population prior to the split of Dmel are maintained through to the split of Dere and Dyak, leading to incomplete lineage sorting and an incongruent genealogy (tree 2). The greater the diversity in the ancestral population and the shorter the time between speciation events, the more likely non-species genealogies are.

Figure 4. Median synonymous trees
Median synonymous branch length trees derived from the genes supporting each of the three trees are drawn to the same scale. The branch spanning the two speciation events is quite short for all trees.

Figure 5. Coalescence probabilities for each tree
Using the formula p(congruence)=1-2/3exp(-t) where t=generations/2Ne, the probability of the species tree (black) and the probability of one of the two alternate trees (grey) was plotted as a function of t.

Figure 6. Clustering of informative sites
The enrichment of informative nucleotide (A) and amino acid (B) substitutions near other substitutions that support the same phylogeny was found for all three trees and is on a scale roughly similar to estimates of linkage disequilibrium. At each informative site in the genome, the counts of informative sites supporting each of the three trees in 1kb windows extending 30kb up and downstream were measured. For each type of informative site, the enrichment of the same type of informative site in each 1kb window was calculated using the observed counts and the expected number of sites based on their genome-wide frequency. Enrichment is $\log_{10}$(observed/expected).

Figure 7. Significance of incongruence
An excess of incongruence above what is expected by chance was observed for the set of all genes (A) as well as the set of genes that consistently supported the same tree across models and species combinations (B). Genes were binned by bootstrap value and the proportion of genes supporting tree 1 (red), tree 2 (green) and tree 3 (purple) were plotted. The expected congruence based on the bootstrap value in each bin (black solid) and the 95% confidence interval based on a X^2 distribution (black dash) demonstrates the excess incongruence.

Figure 8. Sequence & evolutionary gene properties
Sequence and evolutionary properties of the genes are unable to explain the incongruence. Distributions are calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination. The distributions of informative synonymous divergences in genes

supporting each tree reveal a bias toward lower values for the incongruent genes (A). Nearly all genes with little or no informative synonymous divergence, however, are classified as inconsistent (B). Therefore, consistent genes have very similar distributions of informative synonymous divergence across trees (C). Total synonymous divergence is distributed similarly across trees suggesting homoplasy due to increased mutation rates is not causing the incongruence (D). Gene length is slightly higher in tree 1 genes but overall is very similar across trees (E). 3rd codon position GC content is slightly biased toward lower values for Dmel and Dana and higher values for Dere and Dyak, creating a conservative bias for the incongruence (F).

Tables

Table 1. Congruence and fit to data across six models of evolution.

Table 2. Congruence across 21 species combinations.

Supporting Information

Figure S1. Significance of incongruence under six evolutionary models
An excess of incongruence above what is expected by chance was observed for genes from Dmel, Dere, Dyak & Dana using the HKY model (A), the HKY+G model (B), the F3x4 model (C), the F3x4+G model (D), the WAG+F model (E) and the WAG+F+G model (F). Genes were binned by bootstrap value and the proportion of genes supporting tree 1 (red), tree 2 (green) and tree 3 (purple) were plotted. The expected congruence based on the bootstrap value in each bin (black solid) demonstrates the excess incongruence.

Figure S2. Significance of incongruence for 20 species combinations
An excess of incongruence above what is expected by chance was observed using the HKY model for genes from 'Dmel, Dsec, Dsim, Dere, Dyak, Dana & Dpse' (A), 'Dmel, Dsec, Dsim, Dere, Dyak & Dana' (B), 'Dmel, Dsec, Dsim, Dere, Dyak & Dpse' (C), 'Dmel, Dsec, Dere, Dyak, Dana & Dpse' (D), 'Dmel, Dsim, Dere, Dyak, Dana & Dpse' (E), 'Dsec, Dsim, Dere, Dyak, Dana & Dpse' (F), 'Dsec, Dere, Dyak, Dana & Dpse' (G), 'Dmel, Dsim, Dere, Dyak & Dana' (H), 'Dsim, Dere, Dyak, Dana & Dpse' (I), 'Dmel, Dsec, Dere, Dyak & Dana' (J), 'Dsim, Dere, Dyak & Dana' (K), 'Dsec, Dsim, Dere, Dyak & Dana' (L), 'Dsec, Dere, Dyak & Dana' (M), 'Dmel, Dsec, Dere, Dyak & Dpse' (N), 'Dmel, Dsim, Dere, Dyak & Dpse' (O), 'Dmel, Dere, Dyak & Dpse' (P), 'Dsec, Dsim, Dere, Dyak & Dpse' (Q), 'Dsim, Dere, Dyak & Dpse' (R), 'Dsec, Dere, Dyak & Dpse' (S), 'Dmel, Dere, Dyak, Dana & Dpse' (T). Genes were binned by bootstrap value and the proportion of genes supporting tree 1 (red), tree 2 (green) and tree 3 (purple) were plotted. The expected congruence based on the bootstrap value in each bin (black solid) demonstrates the excess incongruence.

Figure S3. Ratio of informative to non-informative synonymous divergence
Although the distributions of the ratio of informative to non-informative synonymous divergence for incongruent genes are biased toward lower values relative to congruent

genes for the set of all genes (A), distributions are similar across trees for the set of consistent genes. Distributions were calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination.

Figure S4. Total synonymous divergence
Total synonymous divergence is distributed similarly across consistent and inconsistent genes (A) as well as across trees for consistent genes (B), with a slight bias toward lower values for inconsistent genes and consistent genes supporting trees 2 and 3. Distributions were calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination.

Figure S5. 1^{st} and 2^{nd} codon position GC content
GC content is distributed nearly identically across species for 1^{st} (A) and 2^{nd} (B) codon positions in all genes. Distributions were calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination.

Figure S6. Sequencing quality scores
Mean sequencing quality scores for coding nucleotides in a gene are distributed nearly identically across trees in the set of all genes for Dere (A), Dyak (B) and Dana (C). Distributions were calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination.

Figure S7. Transition:transversion ratio
Transition:transversion ratios are similarly distributed across trees for the set of all genes. Distributions were calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination.

Figure S8. dN/dS
dN/dS values are similarly distributed across trees for the set of all genes. Distributions were calculated using results from the original maximum likelihood analysis using the F3x4 model and the Dmel, Dere, Dyak & Dana species combination.

Figure S9. Significance of incongruence under RY-coding and F81 model
An excess of incongruence above what is expected by chance was observed for genes from Dmel, Dere, Dyak & Dana using RY-coding and the F81 model. Genes were binned by bootstrap value and the proportion of genes supporting tree 1 (red), tree 2 (green) and tree 3 (purple) were plotted. The expected congruence based on the bootstrap value in each bin (black solid) demonstrates the excess incongruence.

Figure S10. Clustering of informative sites with RY-coding
Controlling for differences in GC content using RY-coding, the enrichment of informative nucleotide substitutions near other substitutions that support the same phylogeny was found for all three trees and is on a scale roughly similar to estimates of linkage disequilibrium. At each informative site in the genome, the counts of informative

sites supporting each of the three trees in 1kb windows extending 30kb up and downstream were measured. For each type of informative site, the enrichment of the same type of informative site in each 1kb window was calculated using the observed counts and the expected number of sites based on their genome-wide frequency. Enrichment is $\log_{10}$(observed/expected).

Table S1. Spearman rank correlations of sequence and evolutionary properties with bootstrap values across sets of genes.

Dsim_Assembly.fasta. Whole genome assembly of *Dsim* in fasta format.

Dsec_Assembly.fasta. Whole genome assembly of *Dsec* in fasta format.

Dere_Assembly.fasta. Whole genome assembly of *Dere* in fasta format.

Dyak_Assembly.fasta. Whole genome assembly of *Dyak* in fasta format.

Dana_Assembly.fasta. Whole genome assembly of *Dana* in fasta format.

Dpse_Assembly.fasta. Whole genome assembly of *Dpse* in fasta format.

Dsim_cDNAs.fasta. cDNA sequences for the set of clear orthologs annotated in *Dsim* in fasta format.

Dsim_translations.fasta. Peptide sequences for the set of clear orthologs annotated in *Dsim* in fasta format.

Dsec_cDNAs.fasta. cDNA sequences for the set of clear orthologs annotated in *Dsec* in fasta format.

Dsec_translations.fasta. Peptide sequences for the set of clear orthologs annotated in *Dsec* in fasta format.

Dere_cDNAs.fasta. cDNA sequences for the set of clear orthologs annotated in *Dere* in fasta format.

Dere_translations.fasta. Peptide sequences for the set of clear orthologs annotated in *Dere* in fasta format.

Dyak_cDNAs.fasta. cDNA sequences for the set of clear orthologs annotated in *Dyak* in fasta format.

Dyak_translations.fasta. Peptide sequences for the set of clear orthologs annotated in *Dyak* in fasta format.

Dana_cDNAs.fasta. cDNA sequences for the set of clear orthologs annotated in *Dana* in fasta format.

Dana_translations.fasta. Peptide sequences for the set of clear orthologs annotated in *Dana* in fasta format.

Dpse_cDNAs.fasta. cDNA sequences for the set of clear orthologs annotated in *Dpse* in fasta format.

Dpse_translations.fasta. Peptide sequences for the set of clear orthologs annotated in *Dpse* in fasta format.

References

1. Russo CA, Takezaki N, Nei M (1995) Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol 12: 391-404.
2. Powell JR (1997) Progress and prospects in evolutionary biology: The Drosophila model: Oxford University Press. 562 p.
3. Lewis RL, Beckenbach AT, Mooers AO (2005) The phylogeny of the subgroups within the melanogaster species group: likelihood tests on COI and COII sequences and a Bayesian estimate of phylogeny. Mol Phylogenet Evol 37: 15-24.
4. O'Grady PM, Kidwell MG (2002) Phylogeny of the subgenus sophophora (Diptera: drosophilidae) based on combined analysis of nuclear and mitochondrial sequences. Mol Phylogenet Evol 22: 442-453.
5. Remsen J, O'Grady P (2002) Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. Mol Phylogenet Evol 24: 249-264.
6. Lemeunier F, Ashburner MA (1976) Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc R Soc Lond B Biol Sci 193: 275-294.
7. Barnes SR, Webb DA, Dover G (1978) The distribution of satellite and main-band DNA components in the melanogaster species subgroup of Drosophila. I. Fractionation of DNA in actinomycin D and distamycin A density gradients. Chromosoma 67: 341-363.
8. Ko WY, David RM, Akashi H (2003) Molecular phylogeny of the Drosophila melanogaster species subgroup. J Mol Evol 57: 562-573.
9. Parsch J (2003) Selective constraints on intron evolution in Drosophila. Genetics 165: 1843-1851.
10. Schlotterer C, Hauser MT, von Haeseler A, Tautz D (1994) Comparative evolutionary analysis of rDNA ITS regions in Drosophila. Mol Biol Evol 11: 513-522.
11. Eisses K (1979) Genetic differntiation within *melanogaster* species group of the genus *Drosophila* (*Sophophora*). Evolution 33: 1063-1068.
12. Solignac M, Monnerot M, Mounolou JC (1986) Mitochondrial DNA evolution in the melanogaster species subgroup of Drosophila. J Mol Evol 23: 31-40.

13. Caccone A, Amato GD, Powell JR (1988) Rates and patterns of scnDNA and mtDNA divergence within the Drosophila melanogaster subgroup. Genetics 118: 671-683.
14. Jeffs PS, Holmes EC, Ashburner M (1994) The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the Drosophila melanogaster species subgroup. Mol Biol Evol 11: 287-304.
15. Shibata H, Yamazaki T (1995) Molecular evolution of the duplicated Amy locus in the Drosophila melanogaster species subgroup: concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation. Genetics 141: 223-236.
16. Nigro L, Solignac M, Sharp PM (1991) Mitochondrial DNA sequence divergence in the Melanogaster and oriental species subgroups of Drosophila. J Mol Evol 33: 156-162.
17. Gailey DA, Ho SK, Ohshima S, Liu JH, Eyassu M, et al. (2000) A phylogeny of the Drosophilidae using the sex-behaviour gene fruitless. Hereditas 133: 81-83.
18. Arhontaki K, Eliopoulos E, Goulielmos G, Kastanis P, Tsakas S, et al. (2002) Functional constraints of the Cu,Zn superoxide dismutase in species of the Drosophila melanogaster subgroup and phylogenetic analysis. J Mol Evol 55: 745-756.
19. Matsuo Y (2000) Molecular evolution of the histone 3 multigene family in the Drosophila melanogaster species subgroup. Mol Phylogenet Evol 16: 339-343.
20. Kopp A, True JR (2002) Phylogeny of the Oriental Drosophila melanogaster species group: a multilocus reconstruction. Syst Biol 51: 786-805.
21. Moriyama EN, Gojobori T (1992) Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics 130: 855-864.
22. Rokas A, Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22: 1337-1344.
23. Gadagkar SR, Rosenberg MS, Kumar S (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zoolog B Mol Dev Evol 304: 64-74.
24. Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51: 588-598.
25. Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating Divergence Times From Molecular Data On Phylogenetic And Population Genetic Timescales. Annu Rev Ecol Syst 33: 707-740.
26. Sanderson MJ, Shafer HB (2002) Troubleshooting molecular phylogenetic analyses. Annu Rev Ecol Syst 33: 49-72.
27. Felsenstein J (2004) Inferring Phylogenies. Sunderland, MA: Sinauer Associates. 664 p.
28. Jermiin LS, Poladian L, Charleston MA (2005) Evolution. Is the "Big Bang" in animal evolution real? Science 310: 1910-1911.
29. Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21: 1455-1458.
30. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54: 743-757.

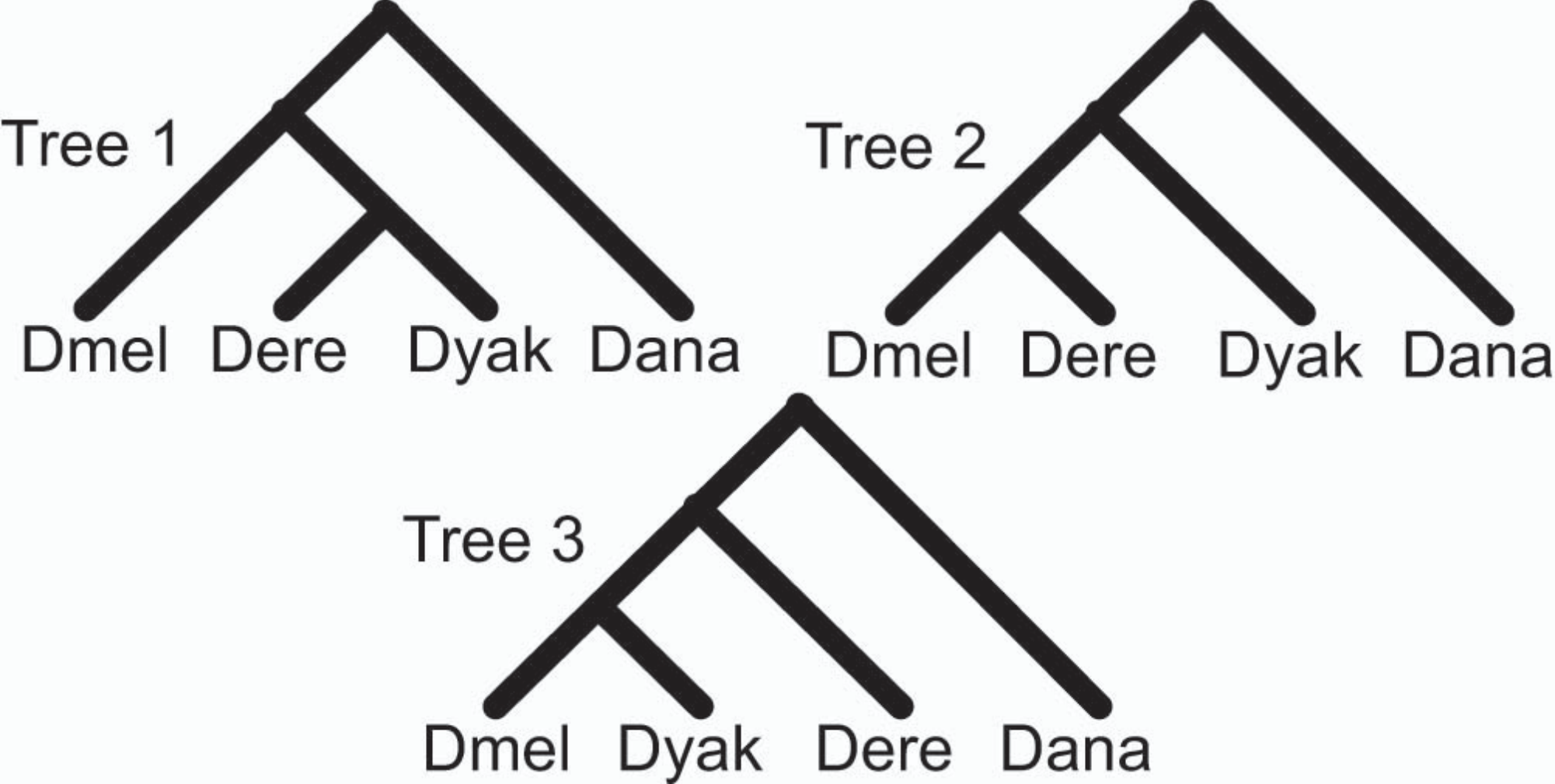
31. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22: 225-231.
32. Andersson JO (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62: 1182-1197.
33. Avise JC, Shapira JF, Daniel SW, Aquadro CF, Lansman RA (1983) Mitochondrial DNA differentiation during the speciation process in Peromyscus. Mol Biol Evol 1: 38-56.
34. Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5: 568-583.
35. Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122: 957-966.
36. Wu CI (1991) Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127: 429-435.
37. Hudson RR (1992) Gene trees, species trees and the segregation of ancestral alleles. Genetics 131: 509-513.
38. Maddison WP (1997) Gene Trees in Species Trees. Syst Biol 46: 523-536.
39. Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, et al. (2000) The population genetics of the origin and divergence of the Drosophila simulans complex species. Genetics 156: 1913-1931.
40. Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68: 444-456.
41. Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. Theor Popul Biol 61: 225-247.
42. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3: 380-390.
43. Rosenberg NA (2003) The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution Int J Org Evolution 57: 1465-1477.
44. Holland BR, Huber KT, Moulton V, Lockhart PJ (2004) Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol 21: 1459-1461.
45. Mossel E, Vigoda E (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309: 2207-2209.
46. Osada N, Wu CI (2005) Inferring the mode of speciation from genomic data: a study of the great apes. Genetics 169: 259-264.
47. Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. Evolution Int J Org Evolution 59: 24-37.
48. Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55: 21-30.
49. Slatkin M, Pollack JL (2006) The concordance of gene trees and species trees at two linked loci. Genetics 172: 1979-1984.
50. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217.
51. Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends in Ecology and Evolution 15: 454-459.

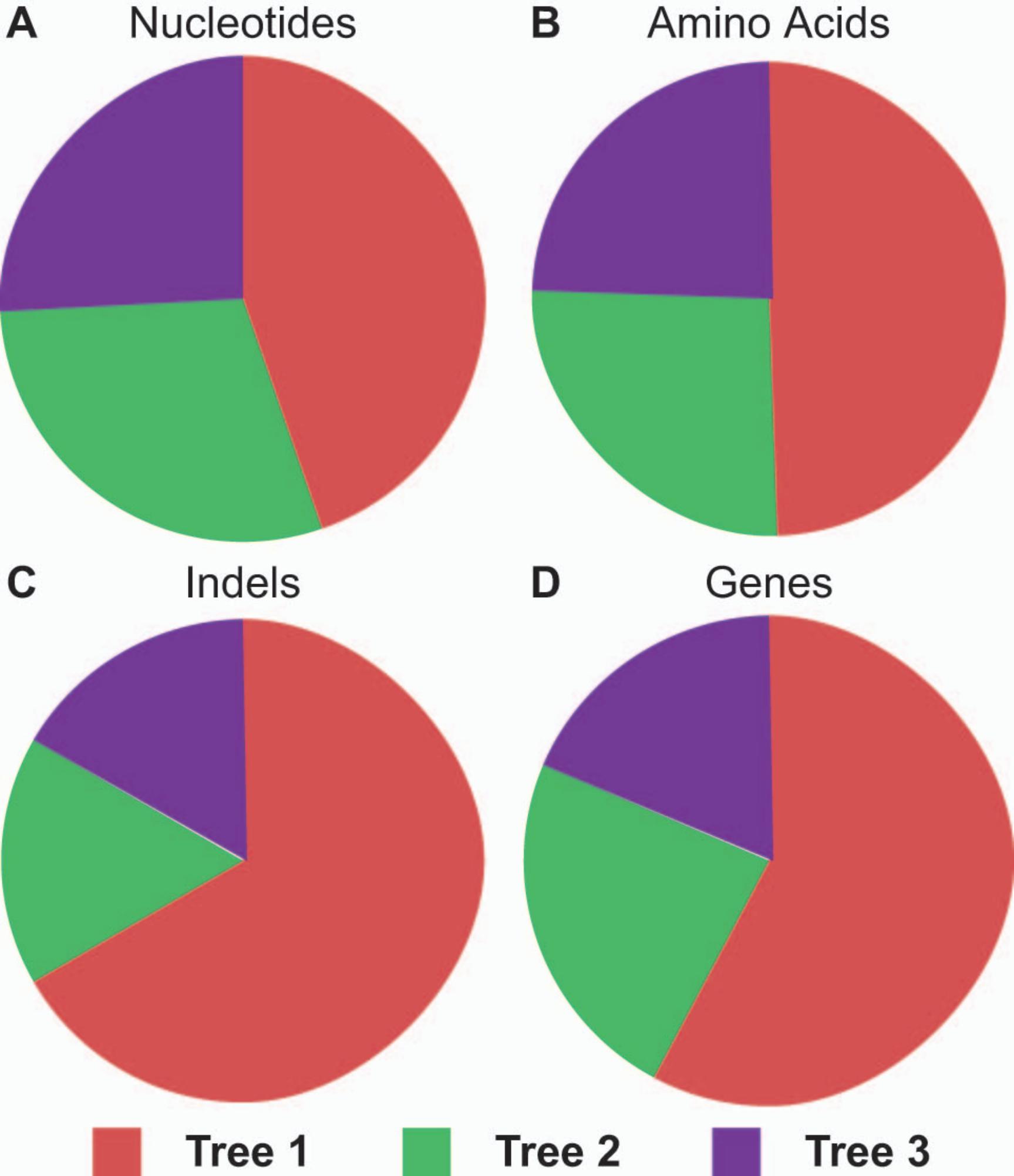
52. Ogurtsov AY, Sunyaev S, Kondrashov AS (2004) Indel-based evolutionary distance and mouse-human divergence. Genome Res 14: 1610-1616.
53. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376.
54. Huelsenbeck JP (1995) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evol 12: 843-849.
55. Holland BR, Jermiin LS, Moulton V (2005) Improved Consensus Network Techniques for Genome-Scale Phylogeny. Mol Biol Evol.
56. Ren F, Tanaka H, Yang Z (2005) An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol 54: 808-818.
57. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.
58. Kingman JFC (1982) The coalescent. Stoch Proc Appl 13: 235-248.
59. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.
60. Nei M (1986) Stochastic errors in DNA evolution and molecular phylogeny. Prog Clin Biol Res 218: 133-147.
61. Sharp PM, Li WH (1989) On the rate of DNA sequence evolution in Drosophila. J Mol Evol 28: 398-402.
62. Li YJ, Satta Y, Takahata N (1999) Paleo-demography of the Drosophila melanogaster subgroup: application of the maximum likelihood method. Genes Genet Syst 74: 117-127.
63. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.
64. Singh RS (1989) Population genetics and evolution of species related to Drosophila melanogaster. Annu Rev Genet 23: 425-453.
65. Wiuf C, Zhao K, Innan H, Nordborg M (2004) The probability and chromosomal extent of trans-specific polymorphism. Genetics 168: 2363-2372.
66. Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics 160: 595-608.
67. True JR, Mercer JM, Laurie CC (1996) Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics 142: 507-523.
68. Takano-Shimizu T (2001) Local changes in GC/AT substitution biases and in crossover frequencies on Drosophila chromosomes. Mol Biol Evol 18: 606-619.
69. Wang W, Thornton K, Emerson JJ, Long M (2004) Nucleotide variation and recombination along the fourth chromosome in Drosophila simulans. Genetics 166: 1783-1794.
70. Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM (2000) Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics 156: 1837-1852.
71. Andolfatto P, Wall JD (2003) Linkage disequilibrium patterns across a recombination gradient in African Drosophila melanogaster. Genetics 165: 1289-1305.
72. Braverman JM, Lazzaro BP, Aguade M, Langley CH (2005) DNA sequence polymorphism and divergence at the erect wing and suppressor of sable loci of Drosophila melanogaster and D. simulans. Genetics 170: 1153-1165.

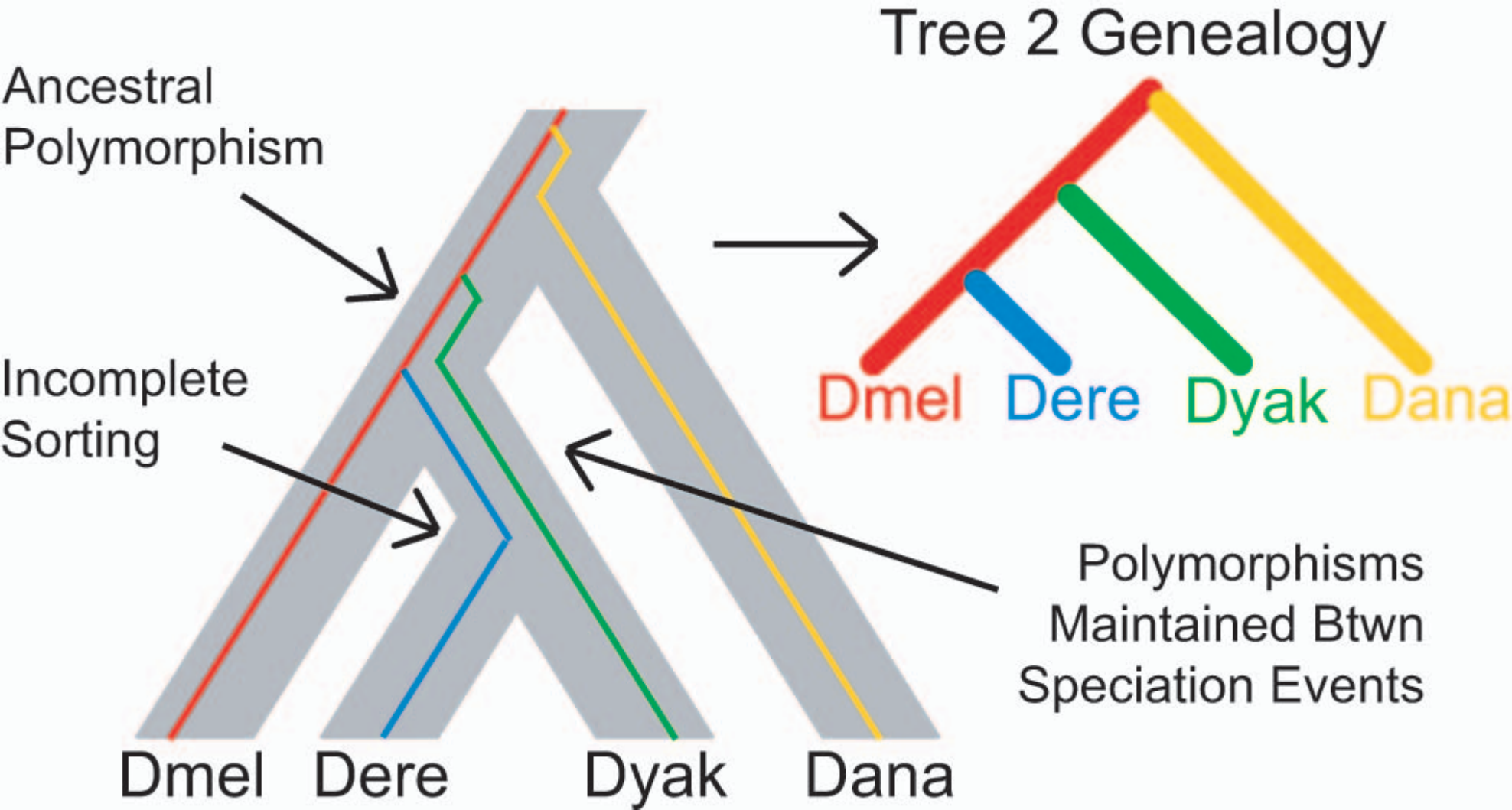
73. Zapata C, Alvarez G (1993) On the detection of nonrandom associations between DNA polymorphisms in natural populations of Drosophila. Mol Biol Evol 10: 823-841.
74. Ohta T, Kimura M (1971) Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68: 571-580.
75. Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172: 1607-1619.
76. Bazin E, Glemin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. Science 312: 570-572.
77. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783-791.
78. Kishino H, Thorne JL, Bruno WJ (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol 18: 352-361.
79. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.
80. Taylor DJ, Piel WH (2004) An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol 21: 1534-1537.
81. Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42: 182-192.
82. Soltis PS, Soltis DE (2003) Applying the bootstrap in phylogeny reconstruction. Stat Sci 18: 256-267.
83. Yang Z (1997) How often do wrong models produce better phylogenies? Mol Biol Evol 14: 105-108.
84. Sullivan J, Swofford DL (2001) Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol 50: 723-729.
85. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr 19: 716-723.
86. Gatesy J, Milinkovitch M, Waddell V, Stanhope M (1999) Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa. Syst Biol 48: 6-20.
87. Gatesy J, Baker RH (2005) Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 54: 483-492.
88. Cunningham CW (1997) Can Three Incongruence Tests Predict When Data Should be Combined? Mol Biol Evol 14: 733-740.
89. Wilcox TP, Garcia de Leon FJ, Hendrickson DA, Hillis DM (2004) Convergence among cave catfishes: long-branch attraction and a Bayesian relative rates test. Mol Phylogenet Evol 31: 1101-1113.
90. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 51: 664-671.
91. Hare M (2001) Prospects for nuclear gene phylogeography. Trends Ecol Evol 16: 700-706.

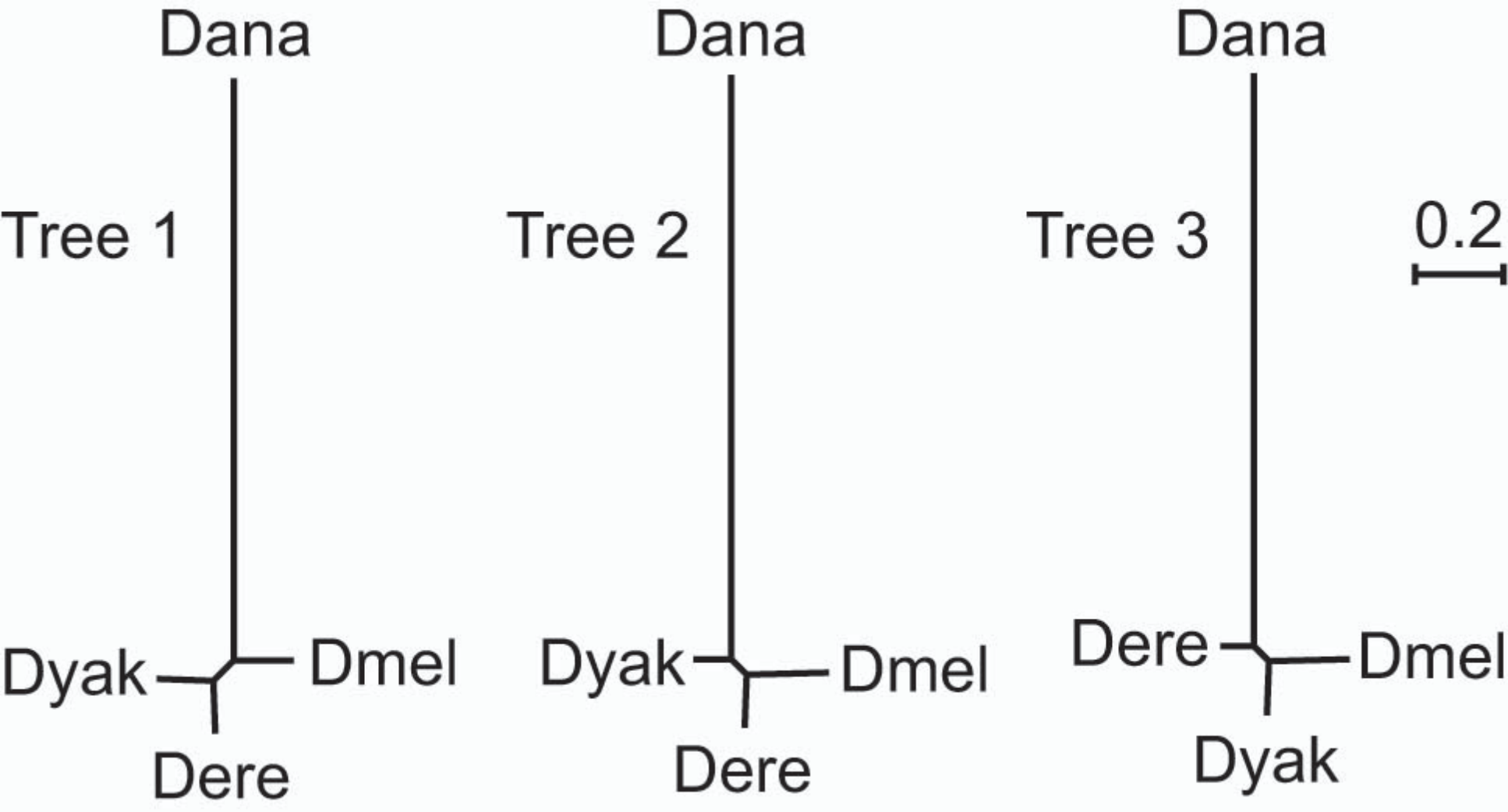
92. Akashi H, Ko WY, Piao S, John A, Goel P, et al. (2006) Molecular evolution in the Drosophila melanogaster species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics 172: 1711-1726.
93. Phillips MJ, Penny D (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28: 171-185.
94. Delsuc F, Phillips MJ, Penny D (2003) Comment on "Hexapod origins: monophyletic or paraphyletic?" Science 301: 1482; author reply 1482.
95. Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92: 11317-11321.
96. Galtier N, Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15: 871-879.
97. Gu X, Li WH (1998) Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc Natl Acad Sci U S A 95: 5899-5905.
98. Husmeier D (2005) Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. Bioinformatics 21 Suppl 2: ii166-ii172.
99. Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Gen 2: 379-384.
100. Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, et al. (2000) The new animal phylogeny: reliability and implications. Proc Natl Acad Sci U S A 97: 4453-4456.
101. Rokas A, Kruger D, Carroll SB (2005) Animal evolution and the molecular signature of radiations compressed in time. Science 310: 1933-1938.
102. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440: 341-345.
103. Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, et al. (2003) The dog genome: survey sequencing and comparative analysis. Science 301: 1898-1903.
104. Nielsen R (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor Popul Biol 53: 143-151.
105. Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution Int J Org Evolution 54: 1839-1854.
106. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158: 885-896.
107. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025-2035.
108. Knowles LL, Maddison WP (2002) Statistical phylogeography. Mol Ecol 11: 2623-2635.
109. Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.

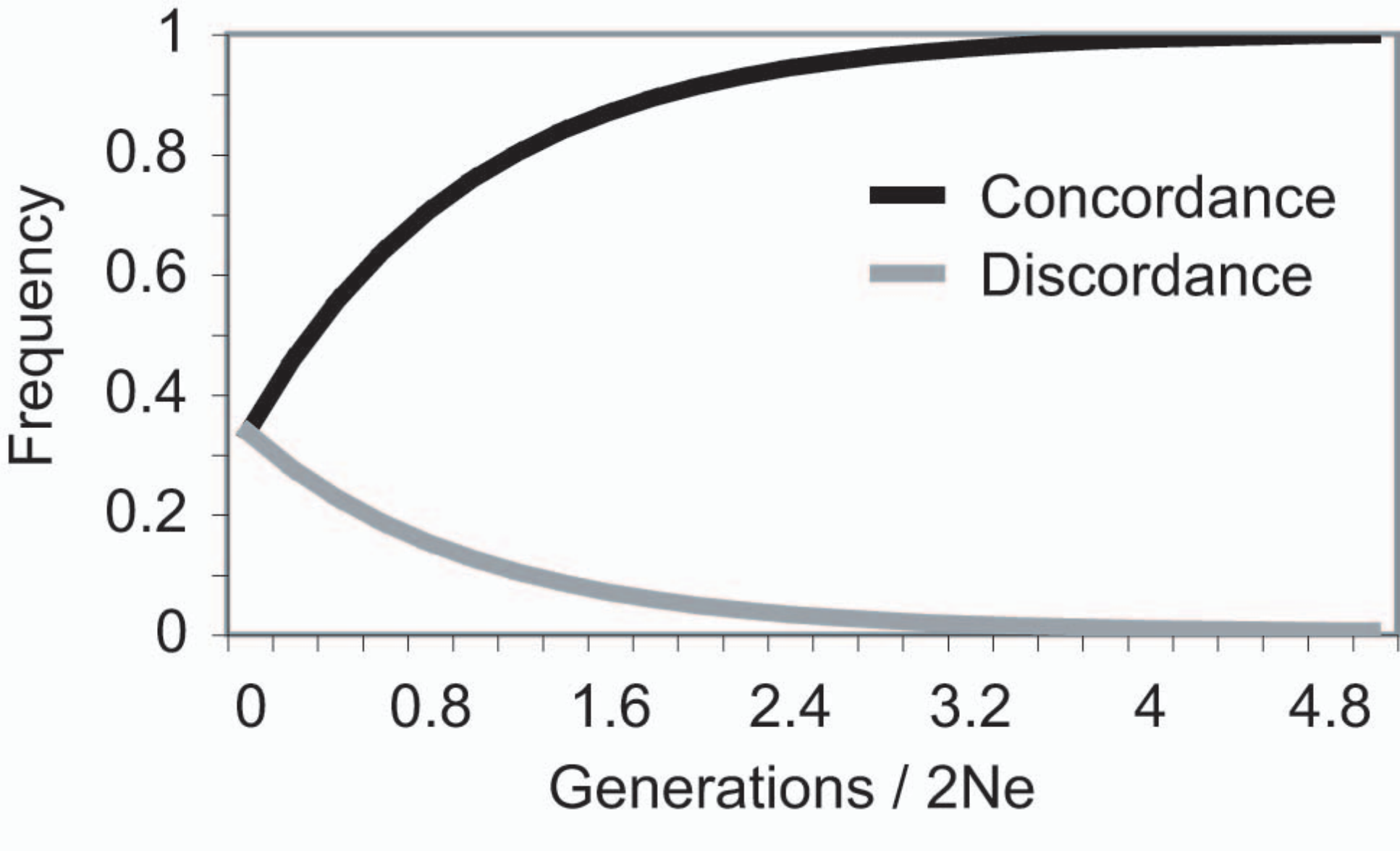
110. Wall JD (2003) Estimating ancestral population sizes and divergence times. Genetics 163: 395-404.
111. Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nat Rev Genet 5: 251-261.
112. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167: 747-760.
113. Felsenstein J (2006) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more Loci? Mol Biol Evol 23: 691-700.
114. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13: 721-731.
115. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708-715.
116. Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19: 1710-1711.
117. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14: 988-995.
118. Birney E, Durbin R (2000) Using GeneWise in the Drosophila annotation experiment. Genome Res 10: 547-548.
119. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39-45.
120. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43.
121. Ihaka R, & Gentleman, R. (1996) R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299-314.

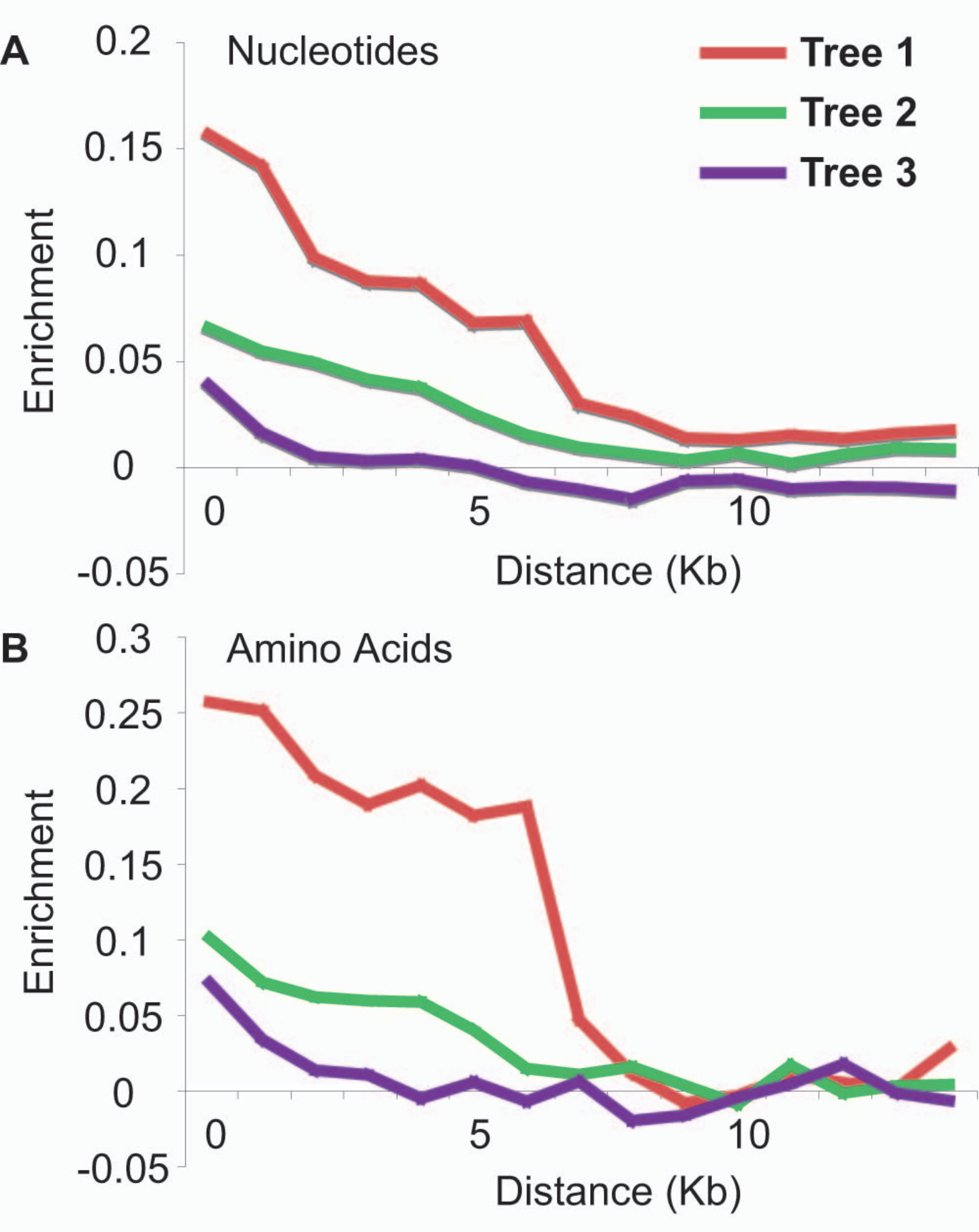

Tree 1
Dmel
Dere
Dyak
Dana
Tree 2
Dmel
Dere
Dyak
Dana
Tree 3
Dmel
Dyak
Dere
Dana

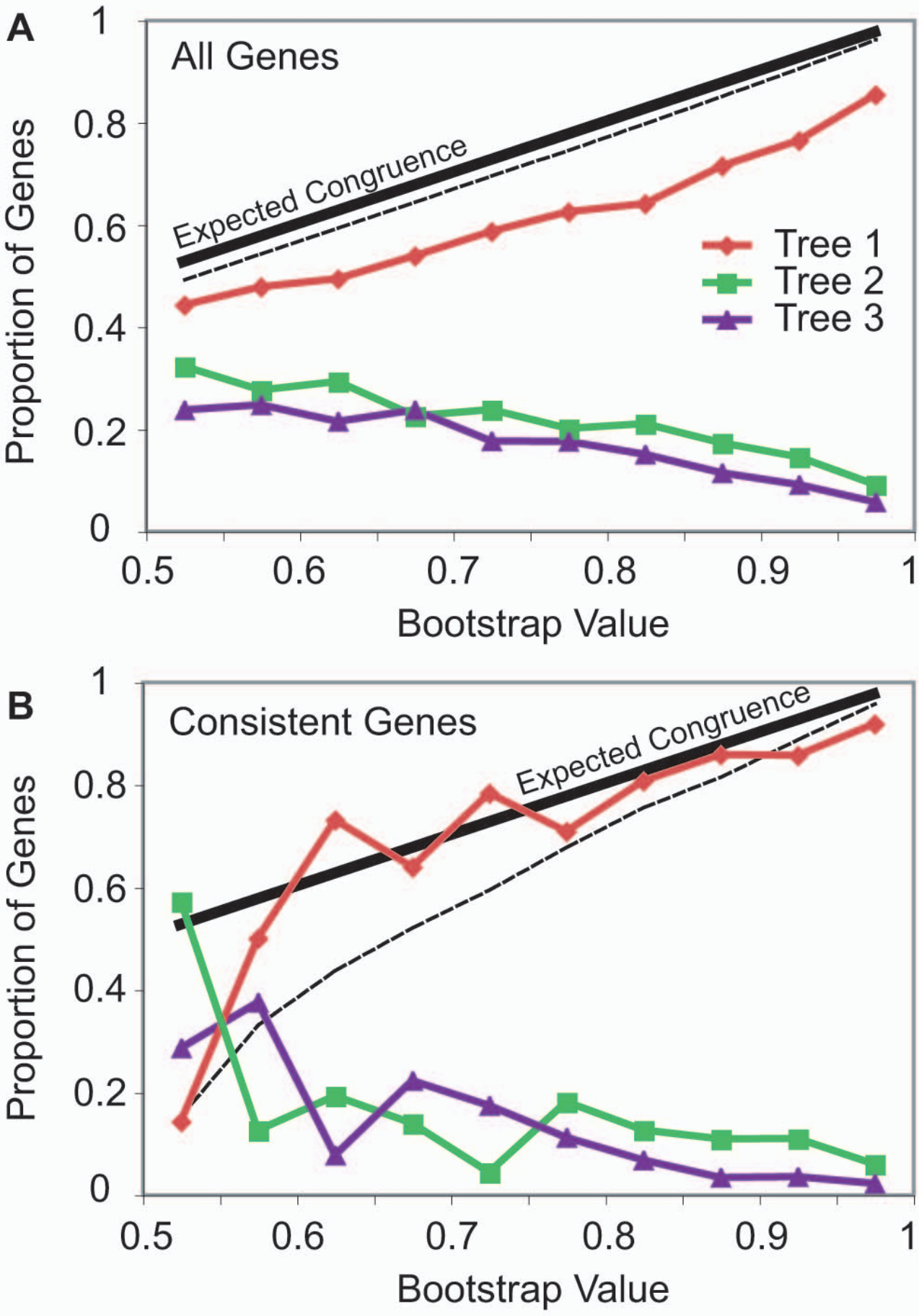

A
Nucleotides
B
Amino Acids
C
Indels
D
Genes
Tree 1
Tree 2
Tree 3

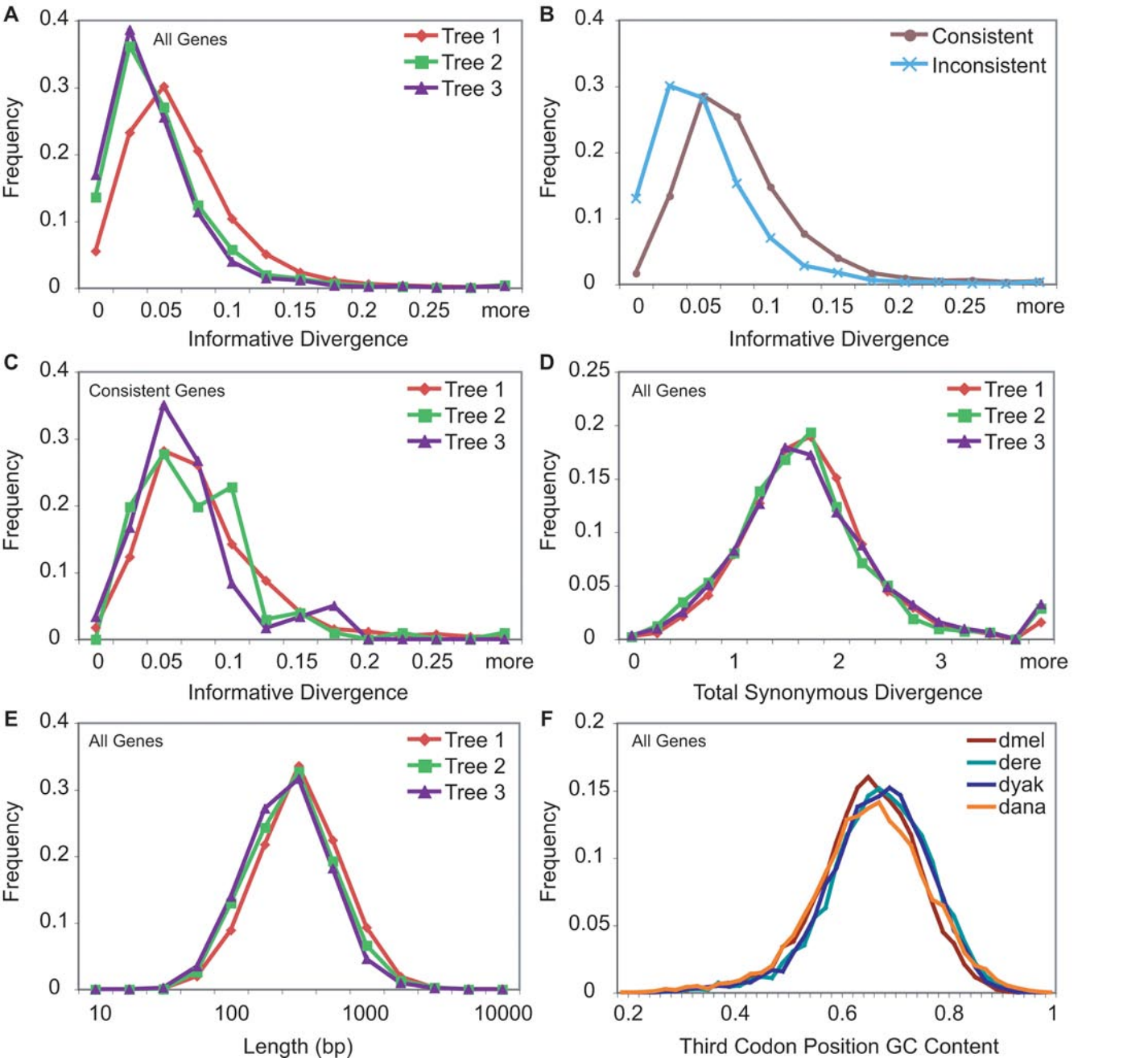

Tree 2 Genealogy
Ancestral
Polymorphism
Incomplete
Sorting
Dmel
Dere
Dyak
Dana
Polymorphisms
Maintained Btwn
Speciation Events
Dmel
Dere
Dyak
Dana


Dana
Tree 1
Dyak
Dmel
Dere
Dana
Tree 2
Dyak
Dmel
Dere
Dana
Tree 3
0.2
Dere
Dmel
Dyak


Frequency
1
0.8
0.6
0.4
0.2
0
Concordance
Discordance
0
0.8
1.6
2.4
3.2
4
4.8
Generations / 2Ne


A
Nucleotides
Tree 1
Tree 2
Tree 3
Enrichment
0.2
0.15
0.1
0.05
0
-0.05
0
5
10
Distance (Kb)
B
Amino Acids
Enrichment
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
5
10
Distance (Kb)


A
All Genes
Expected Congruence
Tree 1
Tree 2
Tree 3
Proportion of Genes
1
0.8
0.6
0.4
0.2
0
0.5
0.6
0.7
0.8
0.9
1
Bootstrap Value
B
Consistent Genes
Expected Congruence
Proportion of Genes
1
0.8
0.6
0.4
0.2
0
0.5
0.6
0.7
0.8
0.9
1
Bootstrap Value


A
All Genes
Tree 1
Tree 2
Tree 3
Frequency
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
more
Informative Divergence
B
Consistent
Inconsistent
Frequency
Informative Divergence
C
Consistent Genes
Tree 1
Tree 2
Tree 3
Frequency
Informative Divergence
D
All Genes
Tree 1
Tree 2
Tree 3
Frequency
0.25
0.2
0.15
0.1
0.05
0
1
2
3
more
Total Synonymous Divergence
E
All Genes
Tree 1
Tree 2
Tree 3
Frequency
10
100
1000
10000
Length (bp)
F
All Genes
dmel
dere
dyak
dana
Frequency
0.2
0.4
0.6
0.8
1
Third Codon Position GC Content

Model	Tree 1	Tree 2	Tree 3	Nuc AIC	AA AIC
HKY	3615 (62.5%)	1284 (22.2%)	885 (15.3%)	0 (0%)	
HKY+G	2882 (49.8%)	1696 (29.3%)	1206 (20.9%)	1 (0.04%)	
F3x4	3215 (56.1%)	1383 (24.1%)	1135 (19.8%)	107 (4.6%)	
F3x4+G	3068 (52.5%)	1455 (25.4%)	1210 (21.1%)	2225 (95.4%)	
WAG+F	2971 (51.9%)	1446 (25.2%)	1312 (22.9%)		618 (26.7%)
WAG+F+G	2917 (50.9%)	1502 (26.2%)	1310 (22.9%)		1694 (73.3%)

Species Combination	Tree 1	Tree 2	Tree 3
dmel/dere/dyak/dana	3615 (62.5%)	1284 (22.2%)	885 (15.3%)
dsim/dere/dyak/dana	3468 (61.0%)	1296 (22.8%)	917 (16.1%)
dsec/dere/dyak/dana	3490 (61.0%)	1359 (23.8%)	869 (15.2%)
dsim/dere/dyak/dana/dpse	3452 (60.8%)	1321 (23.3%)	905 (15.9%)
dsec/dere/dyak/dana/dpse	3447 (60.3%)	1383 (24.2%)	884 (15.5%)
dsim/dsec/dere/dyak/dana/dpse	3419 (60.2%)	1345 (23.7%)	912 (16.1%)
dmel/dere/dyak/dana/dpse	3477 (60.1%)	1371 (23.7%)	935 (16.2%)
dsim/dsec/dere/dyak/dana	3390 (59.7%)	1347 (23.7%)	943 (16.6%)
dmel/dere/dyak/dpse	3365 (58.2%)	1403 (24.3%)	1015 (17.6%)
dsec/dere/dyak/dpse	3324 (58.2%)	1403 (24.6%)	987 (17.3%)
dsim/dere/dyak/dpse	3299 (58.1%)	1374 (24.2%)	1004 (17.7%)
dsim/dsec/dere/dyak/dpse	3249 (57.2%)	1418 (25.0%)	1009 (17.8%)
dmel/dsec/dere/dyak/dana/dpse	3302 (57.1%)	1504 (26.0%)	977 (16.9%)
dmel/dsim/dsec/dere/dyak/dana/dpse	3276 (56.7%)	1506 (26.1%)	996 (17.2%)
dmel/dsim/dere/dyak/dana/dpse	3277 (56.7%)	1502 (26.0%)	1001 (17.3%)
dmel/dsim/dsec/dere/dyak/dana	3240 (56.0%)	1519 (26.3%)	1022 (17.7%)
dmel/dsim/dere/dyak/dana	3229 (55.8%)	1521 (26.3%)	1033 (17.9%)
dmel/dsec/dere/dyak/dana	3215 (55.6%)	1531 (26.5%)	1038 (17.9%)
dmel/dsim/dsec/dere/dyak/dpse	3085 (53.4%)	1578 (27.3%)	1114 (19.3%)
dmel/dsim/dere/dyak/dpse	3084 (53.4%)	1576 (27.3%)	1120 (19.4%)
dmel/dsec/dere/dyak/dpse	3067 (53.0%)	1591 (27.5%)	1125 (19.5%)